RE@CT - IMMERSIVE PRODUCTION AND DELIVERY OF INTERACTIVE 3D CONTENT

Oliver Grau[1], Edmond Boyer[2], Peng Huang[3],

David Knossow[4], Emilio Maggio[5], David Schneider[6]

[1]BBC R&D, London, UK; [2]INRIA, Grenoble, France; [3]Surrey University, Guildford, UK; [4]ARTEFACTO, Rennes, France; [5]Vicon Motion Systems (OMG group), Oxford, UK; [6]Fraunhofer/HHI, Berlin, Germany

E-mail: [1]Oliver.Grau@bbc.co.uk, [2]edmond.boyer@inria.fr , [3]peng.huang@surrey.ac.uk, [4]d.knossow@artefacto.fr, [5]emilio.maggio@vicon.com, [6]David.schneider@hhi.fraunhofer.de

Figure 1 An example of a simple Surface Motion Graph and Animation results.

Abstract: **This paper describes the aims and concepts of the FP7 RE@CT project. Building upon the latest advances in 3D capture and free-viewpoint video RE@CT aims to revolutionise the production of realistic characters and significantly reduce costs by developing an automated process to extract and represent animated characters from actor performance capture in a multiple camera studio.**
The key innovation is the development of methods for analysis and representation of 3D video to allow reuse for real-time interactive animation. This will enable efficient authoring of interactive characters with video quality appearance and motion.

Keywords: Character animation, video game development, immersive media, motion capture.

1 INTRODUCTION

Computer animation is now an essential technique for the production of digital media. Recent advances in graphics hardware have produced video games with a degree of realism only achieved by offline rendered computer generated imagery (CGI) a few years ago. However, applications like games require interactive synthesised animations on the fly. RE@CT [1] aims to revolutionise the production of highly realistic animations of human actors at significantly reduced costs, by developing an automated process which extracts both the visual appearance and motion of actors in a multi-camera studio by combining the latest advances in 3D video into a new character and motion representation.

Technically the production of realistic looking animations of characters is probably the most challenging part of the production of interactive games applications and requires highly skilled experts. The production costs of high-end video games are reaching an average of $10 million for console games and $30k - $300k for casual and social games. Within this budget the production of animations can account for up to 20% of the total budget for smaller and independent production companies without specialist facilities [2]. In these cases animations are often produced with labour-intensive key-frame animation. Since RE@CT aims to automate the animation process to a great extend these costs are expected to be significantly reduced.

The traditional approach is to design the appearance of a character independent from the animation. This is flexible, as the appearance is designed with full artistic freedom on a computer system. The animation is added as a second, separate step using manual animation techniques and motion captured animation data. However, there is currently no seamless process to produce animated content of real people. Applications for this process include production of content with known actors alongside TV- or movie productions (transmedia), capture of heritage or historical events.

RE@CT is developing high-quality capture components based on multiple cameras to allow capture of both the appearance and motion of the actors. This builds upon previously developed techniques to extract 3D information from multiple cameras [4][5]. We extend these capture techniques by active tracking and high-detail face capture and modelling. We also include an immersive feedback system, previously developed for the production of special effects [4] to aid the actors in our capture studio.

Current techniques for the animation of captured content combine manual animation techniques with multi-camera video [3]. RE@CT on the other hand is developing new methods that automatically analyse the captured motion and transform it into a representation that can be used in a games engine to synthesise new movements from the stored data.

The remainder of this paper is structured as follows: The next section gives a brief overview of the project's

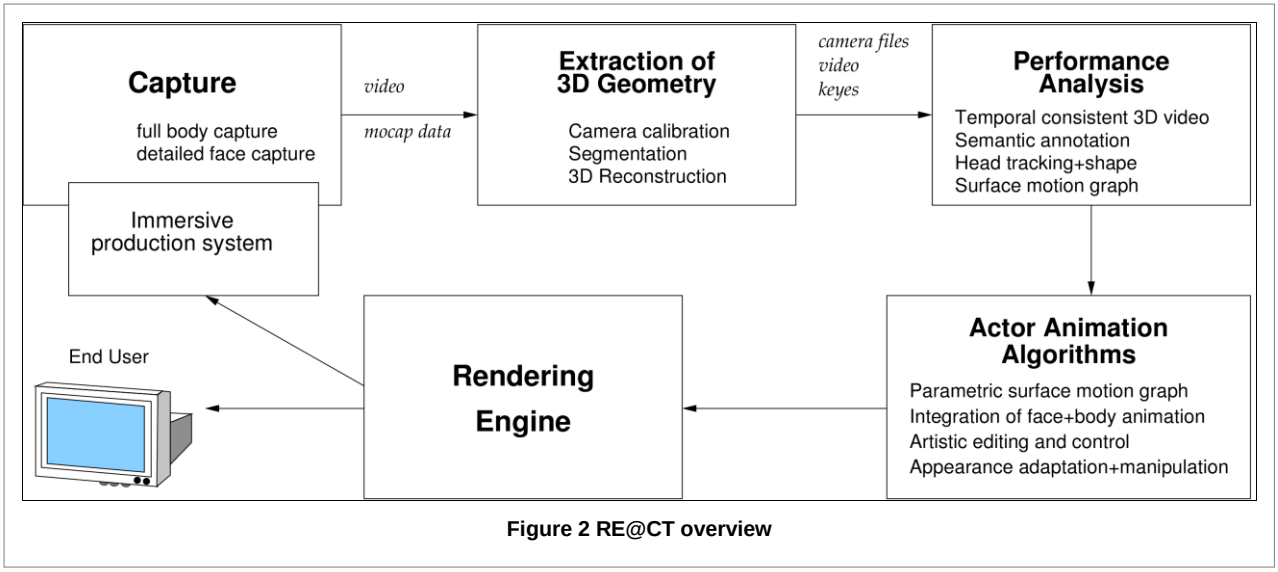

Figure 2 RE@CT overview

components. Section 2 describes the immersive capture system and 3D processing pipeline. In section 4 some details of the performance analysis are given. Section 5 outlines the rendering engine. The paper finishes with first results and conclusions.

2 SYSTEM OVERVIEW

Figure 2 shows the main functional modules of the RE@CT production pipeline. The **capture** module includes the physical studio set-up to capture multiple video streams of an actor. Although the final goal of the project is an image-based system, we include a motion capture (mocap) sub-system for the initial phase of the project and as a reference for verification of the image-based algorithms developed in the project. The video streams are then processed in the **extraction of 3D geometry** module. This module extracts 3D information of the actor on a frame-by-frame basis. Techniques used here are based on visual hull computation and require calibrated cameras and keyed (segmented) images.

The **performance analysis** module performs a temporal alignment of the 3D data. The action is then analysed and a semantic annotation is added. This allows the storage of the action into a surface motion graph.

The **actor animation algorithms** perform a final structuring of the captured data. A parametric representation adds stylistic control and an operator is now able to manipulate the action and to define behaviour of the character as required by the interactive application.

The **rendering engine** is either standalone software or a plugin to a games engine. It allows the play-back and on-the-fly editing of the captured action as controlled by the interactive application. The rendering engine is also used in the **immersive production system** to integrate previously captured action into new capture scenarios.

3 CAPTURE AND EXTRACTION OF 3D GEOMETRY

One of the major aims of RE@CT is to design a studio system to provide high-quality capture of human actions that preserve the full repertoire of body language as well as highly detailed facial expressions used in acting. The project is developing a camera-based acquisition system for whole body and facial 3D video. In order to help actors interact with virtual objects, a novel immersive feedback system will also be investigated.

3.1 Capture

In order to allow for simultaneous development of temporally consistent data representations and actor animation algorithms (see Sections 4 and 5), the project will develop two capture systems: a preliminary hybrid marker-based/marker-less system, and a final video-only markerless system.

3.1.1 Hybrid system

To capture the actors' bodies the hybrid performance capture system will combine video from multiple HD cameras and state-of-the-art marker-based motion capture technology. The marker positions will drive classic articulated models and will provide additional information for 3D reconstruction. Synchronisation between the two systems is achieved by means of industry standard gen-lock signals.

To capture facial expressions in the hybrid system, head mounted capture systems from VICON will be used. Each system is composed of four monochrome cameras positioned as two stereo pairs as showed in Figure 3. The videos are compressed using H.264 and stored on portable logger SSD drives also worn by the actor. The head-mounted cameras synchronize with the rest of the capture system using jam-syncing. An external Time-Code signal is passed to the logger by plugging a cable at the beginning of the capture section. Once the cable is unplugged the logger uses an internal clock to maintain synchronisation with the external system.

3.1.2 Markerless system

The final performance capture system will use multi-view video only and will be composed of a set of commercial HD video cameras with different frame rates: 3CCD lower frame-rate cameras for high quality video, and

single CCD high frame-rate cameras to capture fast motions.

Figure 3 Head mounted camera system used to capture the facial expression of the actors.

In addition to the stationary camera array, a set of multiple pan/tilt/zoom cameras will simultaneously capture high-resolution views of the head. These additional views will provide more detailed information on actors' facial expressions.

3.1.3 System calibration

System calibration in terms of camera locations and lens distortion correction is achieved via standard bundle adjustment techniques. A wand with colour LEDs attached is used as a calibration device. The calibration procedure involves recording views of the wand waved in front of the cameras. Prior knowledge of the LED positioning is used to detect the 2D location of the wand in each frame. Then a bundle adjustment algorithm uses the wand locations to estimate the parameters of the lens distortion model, the camera intrinsic parameters (i.e., focal length, image format, and principal point), and the camera position and orientation.

3.1.4 Face capture

RE@CT features a dedicated capture and processing chain for the actors' heads. This allows cutting to head/face close-ups with a free choice of viewpoint at any time. Also, head shots can be re-animated with the RE@CT animation engine to synthesise new views on the fly as requested by the games engine.

The head capture process must capture data of sufficient quantity and quality for this application. Therefore, footage of the actors' heads is shot on-set by multiple dedicated cameras, recording at full HD or higher quality at a high frame-rate. The cameras have pan/tilt/zoom (PTZ) capability in order to follow the action. They are operated either manually by trained personnel or automatically by robotic camera heads.

The number of head cameras required depends on the number of actors involved in the scene and on the degree of freedom required for image synthesis. In practice, coverage from the front over a baseline angle of 120 degrees is most relevant, allowing the rendering of half-profile views from both directions. An angle of 180 degrees must be covered if full profiles are required. The PTZ cameras must be synchronized as well as calibrated. The RE@CT studio is equipped with PTZ cameras with electronic feedback of zoom settings and an optical calibration system for extrinsic camera parameters, consisting of markers on the studio ceiling which are recorded by a second camera mounted on the actual studio camera.

3.2 3D Processing

3.2.1 Whole body modelling

The 3D geometry of the captured action is computed frame-by-frame using a robust implementation of a visual hull computation. This requires known camera parameters, as computed by the system calibration and a segmentation of the scene into background and foreground, i.e. the actor's silhouette. We use chroma-keying for the segmentation.

3.2.2 Head modelling

To support the computation of 3D information from the PTZ sequences, a set of detailed, fully textured 3D models is generated in advance from the actors involved in the shot. The models are captured with an image-based head-capture rig built by Fraunhofer HHI. The rig comprises multiple stereo pairs of SLR cameras as well as studio flash lighting. The 3D models cover a head-and-shoulder view of the person over an angle of 120 to 180 degrees (frontally). As the reconstruction process is fully image-based the capture process is instantaneous. The actors are captured with different facial expressions, with an emphasis on expressions that have a large impact on the visual hull and overall appearance of the face.

After capture, the footage is analyzed in order to generate temporally and spatially consistent depth information. The captured 3D models support this step by providing detailed information about the head involved. However, pose and expression of the head in the PTZ footage will deviate from the 3D model, which therefore serves primarily as a proxy for analysis.

First, a relationship between each PTZ video stream and the captured model is established. To this end, the individual PTZ video frames are matched with textured renderings of the 3D model using feature matching and image-based optimization techniques. Then the computed relationship of the individual frames to the model is used together with the camera calibration data to establish a consistent spatial relationship between synchronous frames of the different PTZ streams. Finally, the content of each PTZ stream is tracked over time.

The output of the analysis stage is:

depth information for the region of each PTZ video frame showing the actor's head,

inter-view correspondence information, relating the head region of each PTZ frame to its synchronous frames in neighbouring views,

temporal correspondence information, relating the head region of each PTZ frame to the next frame of the same video stream

For motion-graph re-animation of the footage, semantic information on the head footage is required. This will be

obtained by semantically annotating the captured 3D models to describe the facial expressions and propagating this information to the PTZ footage in the registration process. This may require some form of manual intervention for which appropriate tools will be developed.

3.3 Immersive Production System

The processing modules described in the previous section to extract 3D geometry of actors are based on the assumption that the scene can be segmented. Ideally actors are captured in isolation in a controlled studio environment with a chroma-keying facility. That raises the issue that actors have no visual reference: It is very difficult even for trained actors to interact with virtual objects they do not see. The unknown position of another person/object may lead to incorrect pose or gaze of the actor. Similarly, the temporal synchronisation with movements, events, or gestures is difficult to achieve without any feedback.

In order to simplify the capturing process and to obtain better results we will investigate the use of a feedback system previously developed to help the production of special effects [4]. It provides the actor with visual feedback of other virtual humans or objects with which they have to interact. The virtual scene is rendered with accurate timing and projected into the real studio environment without interfering with the capturing. This is achieved with the help of a special keying system, which makes use of retro-reflective cloth. The cameras are equipped with a ring of blue LEDs and the light reflected by the retro-reflective material makes the cloth appear saturated blue as required for chroma-keying. At the same time the actor can observe images from a video projector, as depicted in Figure 4. The light levels have to be balanced so that the projector does not interfere with the chroma-keying, but in practice the set-up is very robust, as the retro-reflective cloth has a high reflectivity peak at a narrow angle of only a few degrees (see [4] for details).

Figure 4: View-dependent projection on retro-reflective cloth

To give the actor an immersive visualisation of the scene, it must be rendered using a view-dependent approach. This requires a rendering system able to render a view of the scene depending on a) the head position of the actor, b) the position of the projector and c) the screen size and position. The position, size and internal projection parameters, i.e. b) and c) are static and can be calibrated and set-up in forehand.

For the actor's head position a real-time head tracking system is required. This is implemented using the video streams of the capture system and real-time processing of the IP-based system. The head position will then be streamed to a rendering module, based on the rendering engine described in section 5. The view-dependent rendering requires a specific setup of the projection.

4 PERFORMANCE ANALYSIS

The system described in the previous section captures shape geometry and appearance independently in each video frame. The performance analysis estimates additional temporal information from this data in order to enable further processing such as actor animation. This additional information includes shape motion that allows a temporally-consistent representation to be built. It also includes semantic information, such as body part labels, to provide more precise knowledge of the observed scene dynamics. This information is used for interaction and animation purposes.

4.1 Temporally Consistent Representations

In order to structure 3D video sequences into meaningful representations for animation, e.g., motion graphs as described in section 5, motion information must be recovered. This key step in the temporal modelling pipeline is implicitly solved when considering traditional motion capture data; however it is a difficult problem when considering 4D sequences of 3D models independently estimated over time. The strategy followed consists of representing 3D video sequences, within a dataset, as motions and deformations of a single known model, i.e. a 3D mesh.

A patch based deformation model [1] is used to track the evolution of a reference mesh over each sequence (see Figure 5). This model preserves the reference mesh consistency over time by enforcing local rigidity constraints. To this end, deformations are encoded as rigid motions of patches. Vertices of the reference mesh are associated to patches at different levels of detail, thereby enabling a coarse to fine strategy when tracking large deformations.

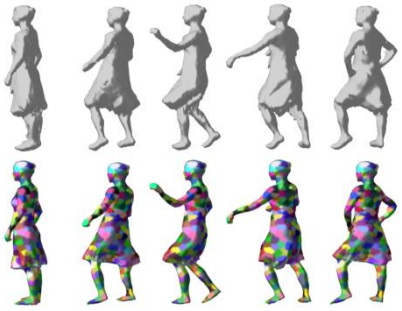

Figure 5: An example of mesh tracking (bottom) given independently estimated 3D models (top).

4.2 Semantic information

Temporally-coherent representations provide dense

motion information for shapes in the form of mesh vertex trajectories over time. These representations do not account for intrinsic properties of shapes neither do they provide compact motion information nor semantic information on shapes, e.g. identification of head, hands, etc. This information however is useful for any shape motion analysis and will enrich the 3D video representations.

When markers are tracked, semantic information can be extracted by fitting a kinematic model to the labelled marker positions using standard marker-based motion capture algorithms. In order to recover similar information in a marker-less environment, 3D video sequences must be analysed. To this purpose, tracked meshes are segmented into parts that exhibit rigid motions over temporal sequences. This is achieved by clustering mesh vertices with respect to their displacement vectors over a time window [10]. Labels can then be associated to rigid parts by either fitting a known articulated model or by considering connectivity information between identified rigid parts.

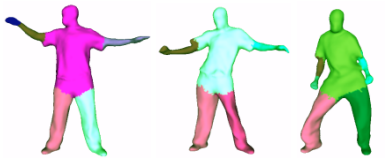

Figure 6: An example of time evolving segmentation of body parts into rigid segments with respect to vertex displacements.

5 ACTOR ANIMATION AND RENDERING

5.1 Actor Animation

The framework for actor animation comprises two stages: pre-processing the database of the 3D video sequences into a Surface Motion Graph [6] or a Parametric Motion Graph [7]; and synthesising actor animation by optimising a path on the graph which best satisfies user input. The user input could be user-defined global constraints (offline) for authoring purposes or interactive control (online) for game-play purposes.

5.1.1 Surface Motion Graph

The Surface Motion Graph consists of a set of motion classes and links between them. Each motion class contains a single motion sequence (e.g. a walk or a hit). Each edge represents a transition across motion classes. The transition is identified as a pair of frames (or a set of overlapped frames), which minimises transition cost from source to target motion. The transition cost is computed as 3D shape dissimilarity between transition frames. A spherical volume-based 3D shape histogram descriptor [1] is used to compare 3D shape dissimilarity between all meshes in the database and pre-compute a 3D shape similarity matrix. Transitions are automatically found and the Surface Motion Graph is constructed. The user is allowed to modify the graph (add/remove some transitions and smooth the transitions by linear/non-linear blending overlapped meshes) to increase the visual quality of final animation.

5.1.2 Parametric Motion Graph

The Parametric Motion Graph can be considered as a natural extension to the Surface Motion Graph. It also consists of a set of motion classes and links between them. Each motion class contains two or more motion sequences (e.g. a slow walk and a fast walk) parameterised to allow generation of any motion in between (e.g. speed of walking). Each link represents a transition between motion classes. A transition may be triggered during the rendering process, for example by a user or by an event in a game. The transition cost is a combination of responsiveness (delay time) and smoothness (3D shape dissimilarity). The 3D shape similarity for all 3D meshes is pre-computed. The dissimilarity for parameterized frames is then approximated as a linear/non-linear interpolation of nearby existing dissimilarity values at run-time to allow real-time transitions between motions.

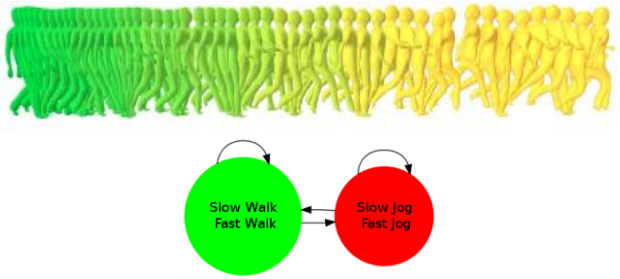

Figure 7: An example of a simple Parametric Motion Graph and Animation results.

5.2 Rendering Engine

5.2.1 Textured Mesh Rendering

View dependant rendering

View-dependent rendering uses a subset of the cameras that are closest to the virtual camera as texture images, with a weight defined according to the cameras' relative distance to the virtual viewpoint. By using the original camera images, the highest-resolution appearance in the representation can be retained, and view-dependent lighting effects such as surface specularity can be incorporated. In practice, the resulting rendering for both original and blended meshes looks realistic and maximally preserves the captured video quality.

Single texture rendering

View-dependent rendering is currently not supported by games engines and other online rendering SDKs. For that reason we will also consider a combined texture map that is compliant with commercially-available rendering SDKs. This approach is common within the context of CG production and will allow the rendering to be done using commercially-available software. Hence, using standard file formats, outputs from the project will be usable within a standard production process.

5.2.2 Interactive rendering tool

Using the technology described in the paragraph above, the rendering engine will be used to interactively animate the meshes. More precisely, an authoring tool will be integrated in the rendering engine to control the motions being rendered in real time. This will allow the end user to control the sequence of motions, defining the motion type (run, walk, jump, etc.), its speed (run slow, fast, etc.) and style.

We aim at proposing a set of very simple interactions for the end-user to control the overall motion and style possibilities. When available, tactile interfaces will be used to ease the motion control.

6 RESULTS

Currently the RE@CT project is focussing on the capture techniques and the integration of the processing modules outlined in previous sections. The first test production will address a cultural heritage application, e.g. for use in museum exhibitions.

RE@CT will provide cultural and historical heritage centres with highly-realistic applications. As an example, RE@CT will allow younger visitors, as well as adults, to control medieval characters and battle in a field using both augmented reality and realism provided by the RE@CT project. Figure 8 shows a typical scenario in such an application. The static assets are themed to a medieval setting in a castle.

Figure 8: Augmented reality application in a cultural heritage setting.

The characters are captured and processed using RE@CT techniques. Figure 9 shows a picture of a recent test production.

Figure 9: Production test in the studio

7 CONCLUSIONS

The RE@CT project aims to provide new tools for the production of interactive character animations. The application scenarios include cultural heritage, education, simulation, and live TV-productions. These applications are often called 'serious gaming'. Furthermore, RE@CT also has the potential to be applied to traditional video game production.

At its core the project is developing tools that capture the action of real actors and then automatically build a data-base from these sampled actions to be used in the interactive context of a games engine to generate new animations on the fly.

This paper has given a brief overview of the individual components that the project is developing. Initial tests have already been carried out to demonstrate the benefits of the techniques. The initial test scenarios are in the domain of cultural heritage. Applications related to broadcast production will follow in the next phase of the project.

References

[1] RE@CT web-site: http://react-project.eu/
[2] The evolution of game animation, white paper, http://www.mixamo.com/c/articles/mixamo_whitepaper
[3] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, Christian Theobalt, Video-based Characters - Creating New Human Performances from a Multi-view Video Database, in ACM Transactions on Graphics 30(4) (Proc. of SIGGRAPH 2011).
[4] O. Grau, T. Pullen, G. Thomas, A combined studio production system for 3D capturing of live action and immersive actor feedback, IEEE Tr. on Systems and Circuits for Video Technology. March 2004.
[5] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In Proc. Of ICCV, pages 915–922, 2003.
[6] P. Huang, J. Starck and A. Hilton. Human Motion Synthesis from 3D Video. In Proceedings of the Twenty-Second IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), pages 1478-1485.
[7] D. Casas, M. Tejera, J-Y Guillemaut and A. Hilton. 4D Parametric Motion Graphs for Interactive Animation. In Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games 2012 (I3D'12).
[8] P. Huang, J. Starck and A. Hilton. Shape Similarity for 3D Video Sequences of People. In International Journal of Computer Vision (IJCV) special issue on 3D Object Retrieval, Volume 89, Issue 2-3, September 2010.
[9] C. Cagniart, E Boyer and S. Ilic. Probabilistic Deformable Surface Tracking from Multiple Videos. In 11th European Conference on Computer Vision, Sep 2010.
[10] R. Arcila, C. Cagniart, F. Hétroy, E. Boyer and F. Dupont. Temporally coherent mesh sequence segmentations. INRIA Research Report RR-7856, 2012.